DOCUMENT RESUME

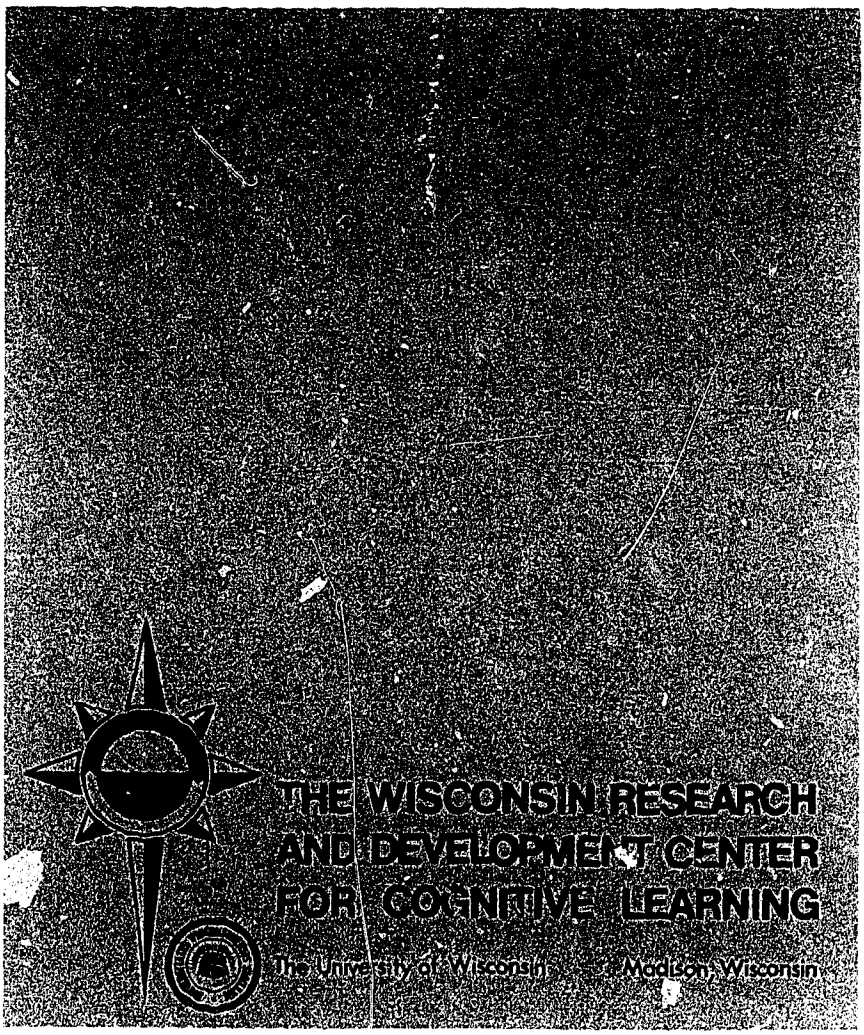
ED 075 491 TM 002 570

AUTHOR Toothaker, Larry E.
TITLE An Empirical Investigation of the Permutation T-Test as Compared to Student's T-Test and the Mann-Whitney U-Test. Report from the Quality Verification Program.
INSTITUTION Wisconsin Univ., Madison. Research and Development Center for Cognitive Learning.
SPONS AGENCY Office of Education (DHEW), Washington, D.C. Research and Development Centers Branch.
REPORT NO TR-174
BUREAU NO BR-5-0216
PUB DATE Feb 72
CONTRACT OEC-5-10-154
NOTE 59p.

EDRS PRICE MF-$0.65 HC-$3.29
DESCRIPTORS *Comparative Analysis; Hypothesis Testing; *Mathematical Models; *Probability Theory; *Statistical Analysis; Technical Reports; *Tests

ABSTRACT

The area investigated in the present study is the comparison of the permutation t-test with Student's t-test and the Mann-Whitney U-test. The comparison was made for small samples for three distributions, including a normal distribution, a uniform distribution, and a skewed distribution. The properties of each test compared were the probability of a Type I error and the power against a location-shift alternative hypothesis. The present research indicates that the permutation t-test is an acceptable statistical procedure for the two-sample problem for the normal and uniform populations and suggests that it might be more desirable than the traditional Student's t-test when sample sizes are proportional to the means and the parent population is nonnormal and asymmetric. Further research is needed before a more definite statement can be made about the permutation t-test when sampling from nonnormal populations. (Author)



TITLE OF DOCUMENT Technical Report No. 174, An Empirical Investigation of the Permutation T-Test as Compared to Student's T-Test and the Mann-Whitney U-Test

AUTHOR(S) Toothaker, Larry E.

Please check the appropriate status of the document:

_____ Copyrighted.

__X__ Not Copyrighted.

If copyrighted, please complete this form.

In order to disseminate as widely as possible timely and significant materials of interest to the educational community, documents abstracted for announcement in the monthly journal of the ERIC system, Research in Education, are made available to users in various ways: in microfiche and/or hard copy sold through the ERIC Document Reproduction Service, and in their original form through the publishers. Credit will be given to the source of each document and if reporduction release is granted, notation will be made of any copyright in the following manner:

"Permission to reproduce this copyrighted material has been granted by_____to ERIC and organizations operating under agreements with the U.S. Office of Education. Further reproduction outside the ERIC system requires permission of the copyright owner."

Please check Level I or II if you grant the ERIC system permission to reproduce the document, or check Level III if permission is not granted.

_____ Level I: in microfiche (4x6 inch film) and hard copy 11x8 1/2 inch photocopy)
_____ Level II: in microfiche (4x6 inch film) only.
_____ Level III: no reproduction release granted.

If you checked Level I or Level II, please sign the following release:

I hereby grant to the Educational Resources Information Center (ERIC) and to organizations operating under agreement with the U.S. Office of Education permission to abstract the above referred to document, to reproduce this document or abstracts thereof by means of microfiche or otherwise, and to disseminate as part of the ERIC system. The authorization to abstract and reproduce copyright materials, however, does not extend to users of the ERIC system.

Signature W. R. Bush per Date 11-28-72
Title Director
Organization Wisconsin Research and Development Center for Cogntive Learning
Address of Organization 1025 W. Johnson Street
Madison, Wisconsin Zip 53706

If you checked Level III, we would like the following information regarding the availability of the document:

Publisher or Distributor:_____

Address:_____

Price per copy_____ Quantity Price_____

Technical Report No. 174

AN EMPIRICAL INVESTIGATION OF THE PERMUTATION T-TEST AS COMPARED TO STUDENT'S T-TEST AND THE MANN-WHITNEY U-TEST

Report from the Quality Verification Program

By Larry E. Toothaker

Mary R. Quilling, Director

Wisconsin Research and Development
Center for Cognitive Learning
The University of Wisconsin
Madison, Wisconsin

February, 1972

This Technical Report is a doctoral dissertation reporting research supported by the Wisconsin Research and Development Center for Cognitive Learning. Since it has been approved by a University Examining Committee, it has not been reviewed by the Center. It is published by the Center as a record of some of the Center's activities and as a service to the student. The bound original is in The University of Wisconsin Memorial Library.

Published by the Wisconsin Research and Development Center for Cognitive Learning, supported in part as a research and development center by funds from the United States Office of Education, Department of Health, Education, and Welfare. The opinions expressed herein do not necessarily reflect the position or policy of the Office of Education and no official endorsement by the Office of Education should be inferred.

Center No. C-03 / Contract OE 5-10-154

STATEMENT OF FOCUS

The Wisconsin Research and Development Center for Cognitive Learning focuses on contributing to a better understanding of cognitive learning by children and youth and to the improvement of related educational practices. The strategy for research and development is comprehensive. It includes basic research to generate new knowledge about the conditions and processes of learning and about the processes of instruction, and the subsequent development of research-based instructional materials, many of which are designed for use by teachers and others for use by students. These materials are tested and refined in school settings. Throughout these operations behavioral scientists, curriculum experts, academic scholars, and school people interact, insuring that the results of Center activities are based soundly on knowledge of subject matter and cognitive learning and that they are applied to the improvement of educational practice.

This Technical Report is from the Quality Verification Program, whose principal function is to identify and invent research and development strategies taking into account current knowledge in the field of statistics, psychometrics and computer technology. The Quality Verification Program collaborates in applying such strategies in research and development. The translation of theory into practice and presentations of exemplars of methodology are challenges which the Quality Verification Program strives to meet.

ACKNOWLEDGMENTS

I wish to thank the members of my thesis committee, Dr. Frank B. Baker, chairman, Dr. Chester W. Harris, Dr. T. Anne Cleary, and Dr. G. William Walster, for their help. I also wish to thank Dr. G. K. Bhattacharya of the Statistics Department for his help.

Appreciation is expressed for computer time given by the Wisconsin Research and Development Center for Cognitive Learning.

I also wish to thank my parents for the support which they have always given me in working toward my goals. My deepest thanks go to my wife, who helped by typing the original thesis and who helped through all my graduate work with her encouragement and sacrifice. Without her, this goal would not have been reached.

TABLE OF CONTENTS

LIST OF TABLES

Abstract

The area investigated in the present study is the comparison of the permutation t-test with Student's t-test and the Mann-Whitney U-test. The comparison was made for small samples for three distributions including a normal distribution, a uniform distribution and a skewed distribution. The properties of each test compared were the probability of a Type I error and the power against a location-shift alternative hypothesis.

The present research indicates that the permutation t-test is an acceptable statistical procedure for the two-sample problem for the normal and uniform populations and suggests that it might be more desirable than the traditional Student's t-test when sample sizes are proportional to the means and the parent population is nonnormal and asymmetric. Further research is needed before a more definite statement can be made about the permutation t-test when sampling from nonnormal populations.

I

INTRODUCTION

A frequently encountered design in educational and psychological research is that which compares some characteristic of two populations. The comparison is usually made by drawing a sample from each of two populations, obtaining a measure of some characteristic of each and testing some function of the measures. If the experimenter desires to test the hypothesis that the population means are equal, then a test statistic commonly used is Student's t-test for two independent samples (Student, 1908). Student's t-test is the statistical procedure chosen most often for the two-sample problem because of a general property of statistical tests: power. The power of a statistical test is the probability of rejecting the null hypothesis given that some alternative hypothesis of interest is true. Another general property affecting the choice of a statistical procedure is the probability of rejecting the null hypothesis falsely, usually known as the probability of a Type I error. The level of the probability of a Type I error is chosen by the experimenter before the experiment takes place. If both populations are normally distributed with equal variances and the alternative hypothesis of interest is that the populations differ only in location, then Student's t-test has the highest power of the available statistical procedures for this situation. Under these conditions, the probability of a Type I error will be exactly the level set by the experimenter.

Thus, if an experimenter is sampling from normal populations with equal variances, and testing a hypothesis of equal population means against a location-shift alternative, Student's t-test is the best statistical test on the basis of power. However, if the populations from which the samples are drawn are not normal, or do not have equal variances, the experimenter might be led to choose a statistical procedure other than Student's t-test. The experimenter would specify the probability of a Type I error and would want to choose the statistical procedure having the highest power for his experimental situation.

A general class of statistical procedures which do not assume normality and which might have high power and an exact probability of a Type I error for non-normal populations are those called distribution-free tests. These tests are not entirely distribution-free because they assume a continuous distribution, although it need not be normal. Two distribution-free tests for the two-sample case are the Mann-Whitney U-test (Mann & Whitney, 1947) and the permutation t-test. The permutation t-test is based upon a distribution obtained by calculating the t-statistic for each permutation of the data. The Mann-Whitney U-test is based upon the ranks of the observations, rather than on the observations themselves. It is of interest to the educational or psychological researcher to know the power of the permutation t-test and the power of the Mann-Whitney U-test against a location-shift alternative for the population with which he is working. Knowing the power and probability of a Type I error of the permutation t-test, the Mann-Whitney U-test and Student's t-test for various populations will allow the experimenter to choose one of the three statistical procedures.

For a normal population it is of interest to know how much power would be lost if the permutation t-test or the Mann-Whitney U-test were used instead of Student's t-test. For a non-normal population, it is of interest to know if the power of the permutation t-test or the Mann-Whitney U-test is larger than the power of Student's t-test. Thus, the populations from which the experimenter could sample might be distributed as the normal, uniform (non-normal but symmetric) and skewed (non-normal and asymmetric) distributions. Knowing the power and probability of a Type I error for the Mann-Whitney U-test, the permutation t-test and Student's t-test for these populations would allow the experimenter to choose one of these three statistical procedures. The present research compares Student's t-test, the Mann-Whitney U-test and the permutation t-test on the probability of a Type I error and the power against a location-shift alternative for the normal, uniform and skewed populations.

The following review of the literature includes a discussion of hypothesis testing in the two-sample case and a detailed discussion of Student's t-test, the Mann-Whitney U-test and the permutation t-test.

Review of the Literature

The two-sample problem is frequently encountered in applied research. Several hypotheses may be made for this design, depending upon the characteristic of the population which the experimenter desires to test. If one desires to test differences between means, the null hypothesis to be tested is that the population means are equal. However, if one desires to test merely that the populations are different, then the null hypothesis to be

tested is that the two independent samples were drawn from the same populations with the same distribution. In the present research, the populations from which the samples were drawn have been specified so the null hypotheses of equal means and equal populations may be considered to be equivalent. The extent to which this equivalence holds is dependent upon the alternative under consideration. The alternative used in the present research was that the populations differed only in location. Thus, the mean of one population was of value μ and the other population, shifted in location by an amount θ, with $\theta > 0$, had a mean of $\mu+\theta$. Thus, only one-tailed tests are considered.

Many statistical procedures have been proposed to test hypotheses of equivalent distributions or hypotheses of equal means. Festinger (1946), Fisher (1925), Kolmogorov (1941), Mann and Whitney (1947), Mood (1950), Pearson (1911), Pitman (1937a), Smirnov (1948), Wald and Wolfowitz (1940), and Wilcoxon (1945) have all given statistical procedures to test the hypothesis of equivalent distributions. Student (1908) presented a statistic whose sampling distribution can be used to test the hypothesis that the means of two normal populations with equal variances are equal.

The statistical procedures included in the present research may be classified on several dimensions. The most obvious classification scheme is by the hypothesis to be tested, which may be classified by terms often used erroneously--parametric and non-parametric. The error which is most often made is that of confusion of the two terms non-parametric (describing the problem) and distribution-free (describing the statistical method used

to solve the problem while making no assumptions about the form of the distribution from which the sample was drawn). Both parametric and non-parametric problems may be solved by statistical methods which may or may not be distribution-free. The Mann-Whitney U-test is used to test the hypthesis of equivalent populations (non-parametric problem) and is a distribution-free statistical procedure. The permutation t-test (or Pitman test) is used to test the hypothesis of equality of means (a parametric problem) and is a distribution-free technique. Student's t-test is used to test the hypothesis of equality of means (a parametric problem) and is not distribution-free. Most distribution-free methods were developed for non-parametric problems and in common usage "non-parametric" is often substituted for "distribution-free."

Another relevant dimension of classification is the assumptions necessary to use the test. One rule accompanying this dimension is that a parametric test in general is more pwerful (i.e., sensitive to change in the factor being tested) than an equivalent non-parametric test <u>if the assumptions for both tests are met</u>. The assumptions may be concerned with the distribution from which the sample was drawn, the independence of the observations or the scale of measurement. It was mentioned above that Student's t-test is parametric. The assumptions for the t-test are: independence of observations, normally distributed errors, equality of variances, and measurement on at least an interval scale. The meaningfulness of the results of the t-test depends upon meeting these assumptions. If a researcher knows that certain of these assumptions cannot be met in his experimental situation, the t-test may not be the appropriate

statistic to be used because another statistic may have higher power than Student's t-test. Most distribution-free tests assume independence of observations and an underlying continuous distribution, but do not make assumptions about the distribution from which the sample was drawn. Parametric tests are generally more powerful than their distribution-free counterparts if their assumptions are met. However, it is logical to question what happens to the statistical test if in fact the assumptions are not met.

The invariance of the probability of a Type I error (α) when the assumptions underlying the test have not been met is known as the robustness of the test (Box and Andersen, 1955). Since parametric tests are most powerful under normal theory assumptions, there is a strong temptation to use these tests when the normality of the distribution is in question. Thus, there has been considerable study of the robustness of parametric tests (Box, 1954a, 1954b, Box and Andersen, 1955) and, correspondingly, there has been considerable study on the power of non-parametric tests. First, the robustness of Student's t-test will be considered and literature pertaining to research done on Student's t-test will be presented. Literature pertaining to the power of the permutation t-test and the Mann-Whitney U-test will follow.

Student's t-test

Student's t-test is used to test the hypothesis of equal population means for the two-sample problem if the populations are normal and have equal variances. Student's t-test is most powerful against a location-shift alternative hypothesis. The test is performed by calculating the two-

independent-sample t-statistic

$$t = \frac{\bar{X} - \bar{Y}}{\sqrt{\frac{\Sigma(X_i - \bar{X})^2 + \Sigma(Y_i - \bar{Y})^2}{m + n - 2}\left(\frac{1}{m} + \frac{1}{n}\right)}}$$

where $\bar{X}$. is the mean of a sample of size m of X_i's and $\bar{Y}$. is the mean of a sample of size n of Y_i's, and determining the probability of obtaining a t-statistic larger than or equal to the original t-statistic by using the tabled t-distribution with m + n - 2 degrees of freedom. If the probability is less than or equal to the probability of a Type I error (usually denoted by α) set by the experimenter, the null hypothesis is rejected. Alternatively, the experimenter may check to see if the calculated t-statistic is greater than or equal to the tabled t-value for the probability of a Type I error and m + n - 2 degrees of freedom. Tables of t are in most elementary statistics texts (see Hays, 1965).

Most research relevant to the robustness of Student's t-test has been done on the one-way analysis of variance, which is the k-sample extension of the two independent sample t-test as introduced by Student (1908). Thus, the analysis of variance research applies to Student's t-test.

Box (1954a) has shown that the one-way analysis of variance, and therefore the two-sample t-test, is robust to violation of the assumption of variance homogeneity if sample sizes are equal. If the sample sizes are unequal, and the variances are also unequal, then the test will have a probability of a Type I error which is smaller than α if the larger sample is from the population which has the larger variance. If the smaller sample came from the population with the larger variance, the test has the probability of a Type I error which is larger than α.

Considering the assumption of a normal population from which the samples were drawn, Kendall and Stuart (1967, p.466) point out that the independence of the numerator and denominator of the t holds only for normal parent populations. If the samples are drawn from a non-normal parent population, the numerator and denominator of the t are not necessarily independent and the dependence affects the probability of a Type I error. However, for large sample size, if the parent population is symmetric or if the samples are of equal size, the t-test is robust to non-normality. Thus the probability of a Type I error is relatively unaffected. Gayen (1941, 1950) found these same results. Srivastava (1958) found that the effect of non-normality on the probability of a Type I error and power of the t-test was not marked if the skewness and kurtosis were small. Little is said of the effect of non-normality of the parent population if the sample size is small for either equal or unequal samples, other than that the t-test should be relatively robust. When sampling from a normal distribution with small samples, the power of the t-test may be calculated exactly (see Milton, 1966). In summary, Student's t-test is relatively robust to violation of assumptions if certain conditions are met. However, in practice it is often difficult to decide if the use of the t-test is likely to be valid or misleading. To aid in deciding on the use of the t-test, preliminary tests have been suggested. The idea of using preliminary tests to determine if the assumptions have been met has been soundly denounced as poor practice (Box and Andersen, 1955) due to the fact that the preliminary test itself then comes under question as to its power with respect to certain factors.

Thus, we would be led to start a long chain of tests each designed to test assumptions for the preceding one. Box and Andersen instead call for tests which are robust and able [illegible] one without preliminary checks on their assumptions.

An alternative to tests which are robust to violations of their distribution assumptions is the derivation of distribution-free statistical procedures which can provide answers to the questions of interest. Such statistical procedures do not assume the observations to be distributed normally, but merely assume that the distribution is continuous. The permutation t-test is such a statistical procedure.

Permutation t-test

The permutation t-test is used to test the hypothesis of equal population means for the two-sample problem if the populations are continuous. The populations do not need to be normally distributed. The permutation t-test is performed by completing the following sequence of events: obtain all possible arrangements (permutations) of the observed data, compute the two independent sample t-statistic for each permutation, arrange the t-statistic in a distribution and determine the probability of obtaining a t-statistic larger than or equal to the original observed t-statistic in this distribution. If the probability is less than or equal to the probability of a Type I error (usually denoted by α) set by the experimenter, the null hypothesis is rejected. Alternatively, the experimenter may check to see if the original t-statistic from the observed data is greater than or

equal the t-statistic which cuts off α-percent of the distribution in the upper tail.

Many of the permutations obtained in the above procedure yield the same statistic. Since it is easier to obtain all possible combinations of m+n divided into m and n, and both procedures yield the same probabilities for the statistic (see Appendix A), the permutation t-test may be based on the distribution of the t-statistic calculated for every possible combination of the observed data. However, the whole procedure depends on the experimenter choosing a probability of a Type I error (α) which divides $\binom{m+n}{m} = (m+n)!/(m!n!)$ evenly.

Permutation tests are based on the fact that any permutation of the observations has an equal chance of occurrence in the distribution of the test statistic. The theoretical basis of the permutation t-test is presented in Scheffé, 1943, pp. 307-308. Simply stated the basis is as follows: the desired property for a statistical procedure which does not assume normality of the population is that the statistical procedure must always yield a region of rejection which has the same probability under the null hypothesis for every possible distribution of measures of interest. Permutation tests guarantee this property because the distribution obtained is based on the data, not on the population, and the probability of the rejection region is always α.

Before the literature on permutation tests can be evaluated, the power of permutation tests must be considered. The power of permutation tests may be generally thought of in two ways: first, as what will be called

an "unconditional power," and second, as a power conditional upon the observations. The conditional power of permutation tests was not used in the present research and is included in the present discussion merely for comparative purposes. There are two types of conditional power of permutation tests: the fixed cut-off point power and a more general power given by Kempthorne (1952). The power used in research by Baker and Collier (1966), Collier and Baker (1966), Kempthorne et al. (1961), and Toothaker (1967) was the conditional power known as the fixed cut-off point power. In the fixed cut-off point procedure the observations are permuted, a specified treatment effect (constant) is added to each observation after the permutation, and the statistic is computed for each permutation. The proportion of permutations with the statistic falling above the fixed cut-off point, usually defined from normal theory for purposes of comparison with normal theory tests, is the conditional power. The fixed cut-off point power is dependent upon the observations. No sampling is done and generalizations may not be made beyond the given set of observations. Also, the fixed cut-off point power is a theoretical power for use primarily in research on the power of permutation tests and is usually not obtained in practice. Another conditional power of permutation tests similar to the fixed cut-off point power is that operationally defined by Kempthorne (1952, p. 219). In the Kempthorne procedure the observations are permuted, a specified treatment effect is added to each observation after the permutation, and the statistic is computed for each permutation. Then for each permutation the statistic is tested via the permutation test: a

permutation distribution of the statistic for observations plus treatment is obtained, the original statistic is compared to this distribution and either an acceptance or a rejection is made. The proportion of the original permutations for which a rejection is made is the power. The conditional power given by Kempthorne is also a theoretical power for use in research on the power of the permutation tests and is not obtained in practice due to the extensive calculations required.

The power of the permutation test which will be called "unconditional power" in the present research is based upon random sampling. The rejection region of the permutation test is conditional upon the observations for each sample, but the power is the proportion of rejections over repeated sampling from some population when the null hypothesis is false. The seemingly illegitimate marriage of a test which was designed to be used on a set of given observations with traditional sampling may be justified as follows: the experimenter usually wants to generalize beyond the set of observations in hand to some population of interest. If the experimenter is going to use the permutation test, and wants to generalize in the usual way to the population from which he has sampled, it is of interest to know the power of the permutation test for repeated sampling from that population. Box and Andersen (1955) point out the difference between unconditional power and conditional power of the permutation test:

> Two alternative views of the nature of the inference in the permutation test can be taken. These differ in the conception of the population of samples from which the observed sample is supposed to have been drawn. On the first view our attention is confined only to that finite population of samples produced by

> rearrangement of observations of the experiment. We prefer to adopt the second view which is that the samples are regarded as being drawn from some hypothetical infinite population in the usual way.

Thus, while the conditional power results from a population dependent upon the observations, the unconditional power is based on random sampling from some population. The obvious advantage of unconditional power is the capability to go beyond the observed data to a population of the statistic based on samples of the given size. The type of power of permutation tests used in the present research is the unconditional power. Thus, the power against the location-shift alternative of the permutation t-test as found in the present research applies to any sample of a given size from a given distribution.

Permutation tests are difficult to perform due to the formidable labor involved in calculating the statistic for all possible permutations, so this procedure was not considered practical until the advent of electronic computers. Because of the lengthy calculations, normal theory tests are used as an approximation for permutation tests even though the rationale for the two types of tests is quite different. The reason the approximation was first suggested was that moment calculations and empirical studies demonstrated the two types of tests to be similar under certain conditions. Most of the literature on permutation tests is on the analysis of variance F-test, and very little is on the permutation t-test. However, results for the one-way analysis of variance are generally applicable to the permutation t-test. Fisher (1935) first introduced the permutation or randomization test as the exact test for testing for differences between means of

two populations when assumptions were not met. Fisher pointed out that the probability of a Type I error for the permutation t-test closely approximated the normal theory probability of a Type I error for the particular problem with which he dealt. Pitman (1937a) was next to consider permutation tests. For the two sample problems, Pitman introduced a test statistic, w, which is a monotonic increasing function of t^2,

$$w \equiv \frac{1}{1 + \frac{N-2}{t^2}} \qquad \text{where } N=m+n, \text{ the combined sample size.} \qquad (2)$$

Pitman (1937b) and Welch (1937) both derived basic results on the permutation test for the analysis of variance for the randomized block and Latin square designs. Both derivations for the analysis of variance held for large sample size and were based on a comparison of moments of the test statistic under normal theory and under permutation. For the randomized block design, Pitman (1937b) and Welch (1937) showed that the F-test may underestimate the significance level if block variances were not equal. However, if the number of blocks is large the underestimation is not serious. Wald and Wolfowitz (1944) derived a general theorem on the limiting distribution of linear forms in the universe of permutations of the observations. They showed that the distribution of the test statistic for the randomized block design is asymptotically the F-distribution underlying normal theory analysis of variance. For Pitman's test, and thus for the permutation t-test, Wald and Wolfowitz showed that the distribution of the test statistic, w, is asymptotically normal. Hoeffding (1952) found that permutation tests for the randomized block design and for the two sample problems are asymptotically as powerful as their related parametric tests. Thus the permutation

test for the randomized block design is asymptotically as powerful as the normal theory F-test, and the permutation t-test is asymptotically as powerful as Student's t-test. Scheffé (1959, Chapter 9) summarized these and other results on permutation tests.

Considerable research has been done on the F-test under permutation in the analysis of variance for various designs, most of it empirical (see Baker and Collier, 1966a; Box and Andersen, 1955; Collier and Baker, 1963; Collier and Baker, 1966; Kempthorne, 1952; Kempthorne et al. 1961; and Toothaker, 1967). The existing research shows that if the assumptions are met and if sample size is not small, the probability of a Type I error and power of the permutation F-test is approximately the same as the power of the normal theory F-test; if the assumptions are not met and if sample size is not small, the probability of a Type I error and power of the F-test under permutation is still fairly close to that of the normal theory F-test, if the violation is not severe.

The study by Box and Andersen (1955) yielded an important result in the study of permutation tests. Box and Andersen introduced a correction for the normal theory F-test. When the degrees of freedom are multiplied by the correction factor, the F-test with the corrected degrees of freedom is an approximate permutation test. The correction factor corrects for the non-normality and heterogeneity of variance of the design. Extensions of this correction procedure have been devised for multivariate situations by Geisser and Greenhouse (1958). The correction factor of Box and Andersen was used in an empirical study by Toothaker (1967) to investigate the

joint effect of variance heterogeneity and block treatment interaction on the F-test under permutation in the randomized block design.

As has been pointed out several times above, most of the research on permutation tests is for large sample size. The present research deals with the comparison of the permutation t-test with Student's t-test and the Mann-Whitney U-test for the normal, uniform and skewed populations for small sample sizes.

The existing literature on the comparison between the permutation t-test and Student's t-test involves the comparison of the power of the two tests. Since Student's t-test is the most powerful test under normal theory, the power of the distribution-free method, the permutation t-test, can be compared to the power of the t-test to measure the loss in power when sampling from a normal distribution. Several measures to compare the power of two tests are available.

One measure to compare the power of two tests is the relative efficiency. The relative efficiency of two tests is defined to be the ratio of the sample sizes necessary to attain the same power against the same alternative, where the sample size in the numerator is that of the most powerful test. Siegel (1956) multiplies the relative efficiency by one hundred and calls it the power efficiency, a more descriptive term. The most commonly used measure is the asymptotic relative efficiency (ARE), defined as the limiting relative efficiency of two tests against a sequence of local alternative hypotheses as the sample size increases. The permutation

t-test has an ARE of 1 when compared to Student's t-test for an alternative of location shift.

A disadvantage of the permutation t-test is that its exact distribution is tedious to enumerate by hand except for very small sample sizes. Also, the distribution of the permutation t-test will be different for every set of actual observations, which are random variables, making it impossible to tabulate the exact permutation distribution of the permutation t-test. With the advent of electronic computers, this disadvantage has become less serious. However, it is still desirable to be able to tabulate the distribution of the statistic for various sample sizes. Rank tests satisfy the desire to be able to tabulate the distribution of the statistic for various sample sizes. The Mann-Whitney U-test is a rank test for the two-sample problem.

Mann-Whitney U-test

One way to remove the variability of the distribution of the test statistic from one set of observations to another is to replace each observation, X_i, with some value, Z_i, for which the permutation distribution of the statistic is the same for every sample of the same size. If these values are chosen to maintain the order relations between two of the values, X_1 and X_2, the ranks of the observations are not the obvious choices. A further desirable aspect of the ranks is that they are invariant under any monotonic transformation of the variable. Therefore, we consider some function of the ranks of the observations. We define the rank of the i^{th}

observation to be its position in the set of the ordered observations, with the smallest receiving the lowest rank. The ordering of the observations, X_i, is one of the N! possible permutations, and the ordering of the ranks, Z_i, is a permutation of the integers one to N. The function of the ranks which has theoretically desirable properties (see Kendall and Stuart, 1967) is the sum of the ranks of one of the samples $R = \sum_{i=1}^{n} Z_i$, where the ranking is done over the total sample. The Mann-Whitney U-statistic is based on such a function. The Wilcoxon and Festinger tests are functions of the U-statistic and thus may be considered equivalent tests.

Rank tests such as the U-statistic are permutation tests. Although few authors point out the fact, many of the rank tests when calculated in their small sample or exact form are permutation tests on the ranks of the observations (see Kruskal and Wallis, 1952, and Wilks, 1962). The rank permutation test exists for not only the two independent sample case but for the two related sample, k independent sample, and k related sample cases. Rank permutation tests also exist for hypotheses of independence (see Hotelling and Pabst, 1936; Kendall and Stuart, 1967; Pitman, 1937a; and Wald and Wolfowitz, 1943). Although only the two independent sample case is considered in this research, future research is planned for the remaining cases.

Specifically, a rank-permutation test exists for the Mann-Whitney U-test. The U-test could be completed for any set of observations by performing a permutation test on the ranks of the observations. The probabilities for possible values of U for a given sample size can be calculated by performing all possible permutations of the ranks for one sample of size

$n(X_1 \ldots\ldots X_n$ where X_i is the i^{th} observation) and the other sample of size $m(Y_1 \ldots\ldots Y_m$ where Y_i is the i^{th} observation) and tabulating the proportion of times a given U value appears.

The Mann-Whitney U-statistic can be defined for a given n and m as the number of times Y rankings exceed X rankings or

$$U = \sum_{i=1}^{n} \sum_{j=1}^{m} h_{ij} \quad \text{where } h_{ij} = \begin{cases} +1 \text{ if } Y_j \; X_i \\ 0 \text{ otherwise} \end{cases} \qquad (3)$$

The calculating formula as given by Mann and Whitney shows the relation between the U statistic and the sum of the ranks:

$$U = nm + \frac{m(m+1)}{2} - R_m \qquad (4)$$

or

$$U = nm + \frac{n(n+1)}{2} - R_n \qquad (5)$$

where R_n is the sum of the ranks of the n observations in the first sample and R_m is the sum of the ranks of the m observations in the second sample. The smaller of (4) and (5) is the tabled value, and the null hypothesis is rejected in favor of the location-shift alternative if values as large or larger than the tabled value are found. Tables now exist for the U-test for probabilities .0005, .005, .0025, .001, .01, .025, .05, and .10 for $m \leq 40$ and $n \leq 20$ (milton, 1964), and the need for exact calculations via permutation does not exist for small samples. For larger sample sizes, the normal approximation is ordinarily used where:

$$E(U) = \frac{nm}{2}$$

and

$$VAR(U) = \frac{nm(n+m+1)}{12}$$

The use of the U-statistic is covered in many elementary statistical tests (see Hays, 1965, and Siegel, 1956) and appears to have heavy usage in all areas of research (see Savage, 1962).

The literature on the Mann-Whitney U-test also includes power comparisons involving the ARE, as was discussed above when comparing the permutation t-test to Student's t-test. As mentioned above, Student's t-test is the most powerful test in the two-sample case if normal theory conditions are met. Therefore, the power of the U-test is necessarily less than that of the t-test for normal theory assumptions. Hodges and Lehmann (1956) have shown that the ARE of the U-test as compared to Student's t-test for a normal distribution is .95 and may never be less than .864 when the location-shift alternative is considered. Hodges and Lehmann also report that the ARE is equal to unity for the uniform distribution. Wetherill (1960) reports that for a gamma distribution with one degree of freedom the ARE is three and for an Edgeworth population with skewness measure $\gamma_1 = .67$ the ARE is unity. So, for non-normal distributions the asymptotic comparison of the power of the U- and t-test shows that the power of the U can be considerably better than that of the t, especially if the distribution is not symmetric.

Small sample power functions for the Mann-Whitney U-test have been derived for several distributions and computations done for at least a few sample sizes. Most of the literature deals with the small sample power of the U-test for the normal distribution and the location-shift alternative. Milton (1966) computed extensive tables of the power of various non-parametric tests against the shift alternative for the normal distribution and offered a direct comparison with the power of Student's t-test. The table of the power of the U-test covers all possible sample size combinations of m,n from 2,1 up to 7,7 for various values of θ. Dixon (1954) and van der Vaart (1950) also have dealt with the power of the U-test and the normal shift alternative. Milton, Dixon and van der Vaart all show that the small sample power of Student's t-test is close to that of the Mann-Whitney U-test for the normal shift alternative. Gibbons (1964), Haynam and Govindarajulu (1966), and Lehmann (1953) have all dealt with the power of the U-test for distributions other than the normal and/or alternative other than location shift. Glazer (1964), Pratt (1964), and van der Vaart (1961) investigated the effect of differences in population variances on the probability of a Type I error of the Mann-Whitney U-test and Student's t-test. The probability of a Type I error of the U-test was less affected by variance differences than the t-test if sample sizes were unequal, but the t-test fared better than the U-test if m=n. Glazer (1964) reported that the small sample power of the t-test was larger than the power of the U-test if m=n or if there were no variance differences. Thus, the U-test is relatively robust to variance differences if m≠n, when compared to the t-test.

Considerable research has been cited above on the power and probability of a Type I error of Student's t-test, the Mann-Whitney U-test, and the permutation t-test. Most of this research has been asymptotic, with exceptions being small sample probability of a Type I error and power of the U-test for selected distributions and alternatives, and the small sample probability of a Type I error and power of the t-test for normal distributions with the shift alternative. There has been essentially no systematic research done on the small sample probability of a Type I error and the power of the permutation t-test for any distribution. The present research investigates empirically the small sample probability of a Type I error and the power of the permutation t-test for normal, uniform and skewed distributions with a location-shift alternative. The probability of a Type I error and power of the Mann-Whitney U-test and Student's t-test are also calculated empirically for comparison purposes and as a check on calculations.

After a general restatement of the problem, Chapter II covers the definition of the power as used in the present study, the procedures for obtaining the power in the computer program used and definitions of the populations.

II

NATURE AND STRUCTURE OF THE PROBLEM

The area investigated in the present study is the comparison of the permutation t-test with Student's t-test and the Mann-Whitney U-test. The comparison was made for small samples for three distributions including a normal distribution, a uniform distribution and a skewed distribution. The properties of each test compared were the probability of a Type I error and the power against a location-shift alternative hypothesis.

Power is generally defined as the probability of a rejection if the alternative is true. More specifically, if X is an observed sample point, ω is the critical region of the test, H_o represents the null hypothesis of equal population means, and H_1 represents the location-shift alternative, then

$$p(X \varepsilon \omega | H_o) = \alpha$$

$$\text{and} \quad p(X \varepsilon \omega | H_1) = 1-\beta = \text{power}$$

where α is the probability of a Type I error and β is the probability of a Type II error. The choice of ω, the critical region, and X, the sample point, depends on the test under consideration. For specific definitions of the power of Student's t-test, the Mann-Whitney U-test and the permutation t-test, the sample point and the critical region must be given in the definition for each. The power of the three statistical procedures in the present study is the unconditional power which is based on random sampling from some population. However, it should be pointed out that the rejection

regions for the Mann-Whitney U-test and Student's t-test are not conditional on the data for any population as is the rejection region for the permutation t-test.

For Student's t-test, a normal theory test, ω is chosen as the top 100α-percent of the theoretical t distribution. The observed sample point, X, is the two-independent sample t-statistic given in (1), above. The test is given by rejecting H_o if t is contained in the rejection region, ω, otherwise failing to reject H_o. The power is then the proportion of rejections over an infinite number of samples and tests of H_o, when the location-shift alternative is true.

For the Mann-Whitney U-test, a permutation test on the ranks of the observations, ω is chosen as the top 100α-percent of the distribution of U obtained by calculating U for each permutation of the ranks of the observations. The observed sample point, X, is the U-statistic for the observed data, and the test is given by rejecting H_o if the U from the observed data is contained in the rejection region, ω, otherwise failing to reject H_o. The power is the proportion of rejections over an infinite number of samples and tests of H_o, when the location-shift alternative is true.

For the permutation t-test, a permutation test on the observations, ω is chosen as the top 100α-percent of the distribution of t obtained by calculating t for each permutation of the observations. The observed sample point, X, is the t-statistic given by formula (1), and the test is

given by rejecting H_o if the t from the observed data is contained in the rejection region, ω, otherwise failing to reject. Then, the power is the proportion of rejections over an infinite number of samples and tests of H_o, when the location-shift alternative is true.

From the above definitions of the unconditional power of Student's t-test, the Mann-Whitney U-test and the permutation t-test, procedures were developed for obtaining estimates of the power and were implemented in the computer program used in the present research. For all three statistical procedures the sampling part of the power procedure was identical and the statistics were all computed on the same observations. A random sample of size n was drawn from a population with mean μ and a second random sample of size m was drawn from a population with mean $\mu+\theta$. Both populations were identical except for the location parameter. For Student's t-test, the t-statistic was computed and the null hypothesis of equal means was rejected if the value of the t-statistic was larger than the tabled 100α-percent value from the t-distribution with $m+n-2$ degrees of freedom. The sampling and computation was done 1000 times and the proportion of rejections yielded an estimate of the power.

For the Mann-Whitney U-test, the same observations as were used for the t-test were ranked and the U-statistic computed on the ranks of one of the samples. The ranks were then permuted and the U-statistic computed for every possible permutation. The original U-statistic was then compared to the distribution of U-values obtained from the permutations and if the original U-statistic was in the 100α-percent rejection region the null

hypothesis of equal means was rejected. The sampling and computation was done 1000 times and the proportion of rejections yielded an estimate of the power.

For the permutation t-test, the t-statistic was computed for the original observations. The observations were then permuted and the t-statistic computed for every possible permutation. The original t-statistic was then compared to the distribution of t-values obtained from the permutations and if the original t-statistic was in the 100α-percent rejection region the null hypothesis of equal means was rejected. The sampling and computation was done 1000 times and the proportion of rejections yielded an estimate of the power.

When θ was equal to zero, the proportion of rejections obtained in the three procedures outlined above was an estimate of the probability of a Type I error for the statistical procedure.

The empirical power and probability of a Type I error for the permutation t-test, Student's t-test and the Mann-Whitney U-test were obtained for normal, uniform and skewed populations. The three distributions of interest were obtained by use of random number generators and a digital computer. To obtain results for the normal population, random samples of size m and n were drawn from the unit normal distribution N(0,1), by use of a random number generator, RANSS (see UWCC User's Manual), and the Control Data 3600 computer. RANSS generates random standard normal deviates by a method which uses pseudo-random odd integers distributed uniformly in the interval $(0,2^{43})$. The uniformly distributed numbers are generated by a power-residue

method (Hull and Dobell, 1962). The procedure uses a starting integer value, X_0, specified by the user, an integer, $a=5^{13}$, and another integer, $m=2^{43}$, called the modulus. A sequence, X_i, of non-negative integers is then defined by the congruence relationship:

$$X_i \equiv 5^{13}X_{i-1} (\text{mod } 2^{43}), \text{ or in general}$$

$$X_i \equiv aX_{i-1} (\text{mod } m) \qquad (6)$$

The method described above is called a power residue method of generating random numbers. The power residue method meets all statistical requirements, i.e., independence of successive values, and numbers distributed as desired as determined by a chi-square test, and it also meets the requirements of a long series of numbers without repetition (see Hull and Dobell, 1962, and IBM, 1959). The power residue method is considered to be satisfactory if it is used correctly (IBM, 1959). A series of numbers produced by a pseudo-random number generator will eventually repeat. Proper use of the power residue method involves choosing the starting value, X_0, the multiplicative constant, a, and the modulus, m, so that they have qualities which yield a long series, X_i. The following limitations, when placed upon the parameters of the congruence relation (6), will yield the longest series of numbers, which will also have good properties statistically:

a) choose $m=2^b$

b) X_0 must be odd and relatively prime to 2^b

c) a must be of the form $a=8c\pm3$, or $a\pm3=8c$ or $c=(a\pm3)/8$ must be an integer.

If the above limitations are placed on the parameters of the congruence relation (6), the generator will produce 2^{b-2} terms before repeating. The RANSS generator has $m=2^{43}$, and $a=5^{13}$ which meet the above requirements because $c=(5^{13}+3)/8$ is an integer. The choice of X_0 odd and relatively prime to 2^{43} will yield a series which has 2^{41} pseudo-random numbers before repeating. Thus, on the order of eight billion numbers may be produced before repeating, which is deemed adequate for the present study.

The random normal generator, RANSS, then uses the X_i values to form a normally distributed random variable. If X_i is the i^{th} variable and $S_n = \sum_{i=1}^{n} X_i$, then $Y = \frac{S_n - n\mu}{\sigma\sqrt{n}}$ is distributed normally with mean=0 and variance = 1, N(0,1), as n approaches infinity due to the Central Limit Theorem (see Mood and Graybill, 1963).

With $n \geq 16$, the approximation of Y to N(0,1) is adequate. Thus, n is taken equal to sixteen, the multiplication and reduction (mod 2^{43}) is repeated sixteen times and the variable Y is returned as the pseudo-random variable distributed N(0,1).

The analysis for the rectangular population was begun by drawing random samples of size m and n from the unit uniform distribution by use of the random number generator RANF (CDC, 1966). RANF generates random numbers in the interval (0,1) by utilizing a power residue method similar to that described above. The parameters of the congruence relation (6) are as follows: $m=2^{47}$ and $a=5^{15}$. The parameters of RANF meet the requirements above if the starting value is an odd integer and relatively prime to 2^{47}. A sequence of non-negative integers is defined by:

$$X_i \equiv 5^{15} X_{i-1} (\text{mod } 2^{47}) \tag{7}$$

which are uniformly distributed in the interval $(0,2^{47})$. To obtain floating point numbers distributed in the interval (0,1), the value of $Y_i=(X_i+1)/2^{47}$ is calculated and returned to the user.

The pseudo-random uniformly distributed numbers returned by RANF were then scaled so that the variance of the population would be unity, the same as the variance of the normal population. The variance of a uniform distribution is given as

$$\sigma^2 = \frac{(a-b)^2}{12} \tag{8}$$

where a and b are the limits of the distribution.

To obtain $\sigma^2=1$, $(a-b)^2$ must equal twelve and a-b must equal the square root of twelve. RANF returns values distributed uniformly in the interval (0,1). If each value returned is multiplied by $\sqrt{12}\approx 3.46$, then the value returned will be distributed uniformly in the interval (0,3.46) and the variance will be approximately one.

The skewed population was derived from a chi-square distribution with three degrees of freedom. The first three moments of the chi-square distribution are v, 2v, and 8v, where v is the degrees of freedom (Kendall and Stuart, 1967, p. 370). The skewness measure

$$\gamma_1 = \sqrt{\frac{2}{\frac{3}{3}}} \tag{9}$$

is then approximately 1.633 for the chi-square with three degrees of freedom. The distribution is unimodal with a positive skew and mean and variance of three and six, respectively.

Since a chi-square variate with N degrees of freedom is defined as

$$X^2_{(N)} = \sum_{i=1}^{N} \left(\frac{Y_i - \mu}{\sigma}\right)^2 = \sum_{i=1}^{N} Z_i^2 \qquad (10)$$

where Z_i is distributed N(0,1), the sum of squares of N unit normal variables is distributed as chi-square with N degrees of freedom. A chi-square variable with three degrees of freedom was generated by calling the unit normal random number generator, RANSS, three times, squaring each unit normal variable, and summing.

The pseudo-random chi-square distributed numbers were then scaled so that the variance of the skewed population would be unity, the same as the variance of the normal population. The variance of a chi-square distribution with three degrees of freedom is six, so each chi-square value was multiplied by $C=1/\sqrt{6}$, yielding a skewed population with mean equal to $3/\sqrt{6}$, and variance equal to one. The skewness measure γ_1 is still equal to 1.633.

The above generation techniques yielded variates distributed as a normal distribution, a uniform distribution and a skewed distribution, respectively.

To obtain results for the probability of a Type I error for the above distributions the values of the probability of a Type I error were chosen for sample sizes such that $\alpha = k/\binom{m+n}{m}$, where k is chosen such that α is close to .05 and $\alpha \leq .05$ if possible. By choosing theoretical values of the probability of a Type I error in this manner, the empirical probability of a Type I error will vary greatly with sample sizes, but will be much

more accurate for a given pair of sample sizes than had the probability of a Type I error been chosen such that $\alpha \leq .05$ for all sample sizes. Also, certain values of the sample sizes, such as samples of sizes two and three, could not possibly yield values of theoretical probability of a Type I error less than .05 and would have had to have been left out of the study. Such choice of the theoretical probability of a Type I error also made the specification of the power considerably easier, since the exact probability of a Type I error and thus the exact critical value could be obtained.

To obtain the results for the power for the three statistical procedures, $\theta > 0$ was defined such that the levels of power of Student's t-test would be .30, .60, and .90 for the normal distribution. The defined θ was used for all three statistical procedures and for all three distributions.

Specification of θ for the normal distribution was made through the definition of the non-centrality parameter, δ^2, for the non-central t-distribution as given by Scheffé (1959, p. 41),

$$\sigma^2\delta^2 = \psi' B^{-1} \psi \qquad (11)$$

where ψ is the column vector of contrasts on the cell means, μ_1 and μ_2, and $B = \frac{1}{\sigma^2} \ddagger$ where $\ddagger$ is the variance of the desired contrast. Since the t-test deals with the difference between means, the contrast desired is $\mu_1 - \mu_2$,

so

$$\psi = \psi' = (\mu_1 - \mu_2)$$

$$\ddagger = \sigma^2 \left(\frac{1}{m} + \frac{1}{n}\right)$$

then $$B = \left(\frac{1}{m} + \frac{1}{n}\right)$$

$$B^{-1} = \left(\frac{1}{m} + \frac{1}{n}\right)^{-1}$$

and $$\sigma^2\delta^2 = (\mu_1-\mu_2)\left(\frac{1}{m} + \frac{1}{n}\right)^{-1}(\mu_1-\mu_2)$$

or $$\sigma^2\delta^2 = \frac{(\mu_1-\mu_2)^2}{\left(\frac{1}{m} + \frac{1}{n}\right)} \tag{12}$$

Setting $\sigma^2=1$, and solving for $\mu_1-\mu_2$ in terms of δ yields

$$\mu_1-\mu_2 = \delta\sqrt{\frac{1}{m} + \frac{1}{n}} = \theta \tag{13}$$

Several FORTRAN subprograms were used to obtain the values of θ, which were utilized in the main program. First, the exact t-value was obtained for the exact probability of a Type I error for given sample sizes through use of a subprogram written to compute exact probabilities for the F-distribution (see Baker and Collier, 1966b). The obtained t-value and the desired probability of a Type II error (1-desired power value) were used in another subprogram written by Milton (see UWCC User's Manual under "New Subprograms") to yield the appropriate non-centrality parameter, δ, for those sample sizes. Given δ, m and n, the value of θ was computed. The power results could be obtained by drawing one of the samples from a

distribution with mean $\mu + \theta$ and the other from a distribution with mean μ. One method for achieving the desired result would be to alter the random number generators. However, it was not necessary to alter the random number generators to sample from populations with means $\mu + \theta$ for the following reason: if a constant θ is added to every score in a distribution with mean μ, the mean of the new distribution is simply $\mu + \theta$. Thus, by sampling from a distribution with mean μ and adding the defined θ to each value obtained, the result is the same as if sampling had been done from a distribution with mean $\mu + \theta$. For the normal population, $\theta + \mu_1$ because $\mu_2=0$. The samples drawn for the power results were as if they had been drawn from the normal distributions N(0,1) and n(θ,1), from the rectangular distribution $f_2(x)$ and $f_1(x+\theta)$ and from the skewed distribution $f_2(x)$ and $f_1(x+\theta)$, where θ is defined as in (13) above. The values of θ for all sample sizes considered in the present study are given in Appendix B.

The sample sizes considered in the present research are the nine arrangements of (2,3), (2,4), (2,5), (3,3), (3,4), (3,5), (4,4), (4,5) and (5,5). These sample sizes were part of a larger set originally chosen because of existing exact probabilities of the Mann-Whitney U-statistic in table form. Consideration of computing time and programming difficulty then narrowed the range of sample sizes to the above set.

The empirical small sample power and size for the permutation t-test, Student's t-test and the Mann-Whitney U-test were obtained by means of a computer program written for this purpose by the author. The program MONTE1 was written in FORTRAN and was run on the Control Data Corporation

3600 computer. For a given sample size the program is designed to draw samples from the appropriate population, add the specified θ (null or non-null) to the data, complete the permutation procedure for the U-test and the permutation t-test, complete the normal theory test for Student's t-test by use of the appropriate value from the t-distribution. The program is designed to then repeat the entire procedure 1000 times. The number of samples to be drawn was determined strictly by consideration of the computing time involved. The number 1000 was the largest possible number of samples which could be analyzed without using an inordinate amount of computer time. After the 1000 samples have been drawn, the program is designed to then print out the estimated probability of a Type I error and power of each of the three tests for the given sample size. In addition, it was thought advisable to check for influence of the size of the sample to which the θ was added, so two sets of power values are printed, one set for θ being added to the larger of the two samples and one set for θ being added to the smaller of the two samples.

III

RESULTS

Probability of a Type I Error

The empirical values of the probability of a Type I error of Student's t-test, the Mann-Whitney U-test, and the permutation t-test for various small sample sizes from the previously specified normal, uniform, and skewed populations are given in Table 1 as evidence verifying the Monte Carlo procedures. The theoretical probability of a Type I error is given as α. Only two empirical values of the probability of a Type I error in Table 1 were larger than that expected from sampling variability. For sample size (4,4) from the skewed population, the values of .044 and .044 for the Mann-Whitney U-test and the permutation t-test were more than $2\sigma_p$ larger than .0286, the theoretical α. For equal sample sizes from the skewed population, there was a trend of empirical values of the probability of a Type I error for the Mann-Whitney U-test and the permutation t-test which were larger than both the theoretical α and the value for Student's t-test. Also, for unequal sample sizes from the skewed population, the values for the Mann-Whitney U-test and the permutation t-test followed the opposite trend; that is, they were less than the theoretical α in four of the six cases and less than the value for Student's t-test in five of the six cases.

The remaining values of the empirical probability of a Type I error were within the bounds of sampling variation, and there were no other

TABLE 1

The Empirical Probability of a Type I Error for Three Two-Sample Statistics, for Three Parent Populations, and for Various Sample Sizes (the values in the table are the proportion of rejections in 1,000 random samples)

Sample Sizes and σ_p Values	α	Statistical Test	Normal	Uniform	Skewed
(2,3)	.10	Student's t	.105	.109	.104
		Mann-Whitney U	.109	.105	.094
σ_p = .0095		Permutation t	.109	.105	.094
(3,3)	.05	Student's t	.055	.056	.046
		Mann-Whitney U	.056	.051	.053
σ_p = .0069		Permutation t	.056	.051	.053
(2,4)	.0667	Student's t	.079	.072	.073
		Mann-Whitney U	.074	.065	.059
σ_p = .0079		Permutation t	.074	.065	.059
(3,4)	.0286	Student's t	.034	.037	.033
		Mann-Whitney U	.031	.027	.028
σ_p = .0053		Permutation t	.031	.027	.028
(4,4)	.0286	Student's t	.029	.025	.039
		Mann-Whitney U	.029	.021	.044[a]
σ_p = .0053		Permutation t	.029	.021	.044[a]
(2,5)	.0476	Student's t	.046	.043	.055
		Mann-Whitney U	.049	.050	.043
σ_p = .0067		Permutation t	.049	.050	.043
(3,5)	.0357	Student's t	.038	.034	.047
		Mann-Whitney U	.039	.036	.041
σ_p = .0059		Permutation t	.039	.036	.041
(4,5)	.0317	Student's t	.026	.037	.035
		Mann-Whitney U	.026	.033	.040
σ_p = .0055		Permutation t	.024	.032	.039
(5,5)	.0476	Student's t	.055	.054	.045
		Mann-Whitney U	.055	.050	.050
σ_p = .0067		Permutation t	.055	.052	.050

[a]The observed empirical probability is more than $2\sigma_p$ from α.

consistent trends evident in the empirical values. The equality of the empirical values in Table 1 for the Mann-Whitney U-test and the permutation t-test for all sample sizes other than (4,5) and (5,5) is due to the fact that when the number of combinations is small, the rejection region for both tests contains a very small number of points. Thus, only a few combinations of the data result in a rejection with the permutation t-test, and the same exact combinations are the ones which yield rank sums large enough to cause a rejection with the U-test. When the sample size gets larger, such as (4,5) and (5,5), there are more points in the rejection region, therefore the chance of a combination of the data to reject on one test and not on the other.

Power

The values of the empirical power of Student's t-test, the Mann-Whitney U-test, and the permutation t-test for various small sample sizes from the previously specified normal, uniform, and skewed populations are presented in Table 2. As was the case with the probability of a Type I error, the values of empirical power for the permutation t-test and the Mann-Whitney U-test are identical within each population for sample sizes smaller than (4,5).

For the normal and uniform populations, the power of Student's t-test was generally larger than the power of the permutation t-test for both the "small" and "large sample addition procedure" and for all sample sizes. Of the 108 cases available (three levels of θ, nine sample sizes for the large and small sample addition procedures for each of two populations) there were 102 cases where the power of Student's t-test was larger than that of the permutation t-test and 37 which were larger

TABLE 2

The Empirical Power for Three Two-Sample Statistics, for Various Values of θ [Small (S), Medium (M), and Large (L)] With θ Added for Either the Small Sample and for Samples of Various Sizes (The values in the table are the proportions of rejections in 1,000 samples.)

Normal Population

Sample Size	θ	Small Sample Addition Procedure			Large Sample Addition Procedure		
		t	U	Permutation t	t	U	Permutation t
(2,3)	S	.259	.258	.258	.270	.255	.255
	M	.600	.588	.588	.599	.584	.584
	L	.893	.882	.882	.901	.886	.886
(3,3)	S	.293	.288	.288	.282	.277	.277[a]
	M	.596	.574	.574	.601	.569[a]	.569[a]
	L	.902	.881[a]	.881[a]	.894	.877	.877
(2,4)	S	.287	.283	.283	.309	.303	.303
	M	.598	.573	.573	.610	.591	.591
	L	.886	.871	.871	.897	.884	.884
(3,4)	S	.311	.302	.302	.302	.295	.295[a]
	M	.612	.568[a]	.568[a]	.615	.569[a]	.569[a]
	L	.904	.881[a]	.881[a]	.892	.864[a]	.864[a]
(4,4)	S	.307	.290	.290	.277	.280	.280
	M	.615	.599	.599	.590	.569	.569[a]
	L	.918	.893[a]	.893[a]	.894	.874[a]	.874[a]
(2,5)	S	.295	.297	.297	.288	.255[a]	.255[a]
	M	.607	.575[a]	.575[a]	.597	.565[a]	.565[a]
	L	.916	.871[a]	.871[a]	.913	.873[a]	.873[a]
(3,5)	S	.313	.309	.309	.301	.285	.285
	M	.610	.580	.580	.576	.560	.560
	L	.889	.872	.872	.890	.863[a]	.863[a]
(4,5)	S	.302	.276	.292	.312	.293	.304
	M	.611	.577[a]	.600	.577	.556	.568
	L	.901	.890	.898	.893	.867[a]	.878
(5,5)	S	.312	.293	.309	.294	.287[a]	.294
	M	.601	.576	.606	.583	.548	.581
	L	.908	.886[a]	.905	.892	.874	.884

[a]The observed proportion is more than $2\sigma_p$ from the observed proportion for t, where, for S, M, and L, σ_p = .0145, .0155, and .0095, respectively.

TABLE 2 (Continued)

Uniform Population

Sample Size	θ	Small Sample Addition Procedure t	Small Sample Addition Procedure U	Small Sample Addition Procedure Permutation t	Large Sample Addition Procedure t	Large Sample Addition Procedure U	Large Sample Addition Procedure Permutation t
(2,3)	S	.263	.252	.252	.260	.250[a]	.250[a]
	M	.567	.554	.554	.574	.542[a]	.542[a]
	L	.906	.897	.897	.919	.898[a]	.898[a]
(3,3)	S	.276	.265	.265[a]	.257	.239[a]	.239[a]
	M	.555	.514[a]	.514[a]	.537	.499[a]	.499[a]
	L	.919	.915	.915	.927	.928	.928
(2,4)	S	.279	.274	.274	.270	.266[a]	.266[a]
	M	.541	.513	.513	.553	.512[a]	.512[a]
	L	.922	.905	.905	.906	.892	.892
(3,4)	S	.255	.234[a]	.234[a]	.255	.217[a]	.217[a]
	M	.546	.496[a]	.496[a]	.570	.495[a]	.495[a]
	L	.919	.924	.924	.939	.928	.928
(4,4)	S	.252	.214[a]	.214[a]	.253	.235[a]	.235[a]
	M	.561	.511[a]	.511[a]	.562	.515[a]	.515[a]
	L	.928	.900[a]	.900[a]	.930	.902[a]	.902
(2,5)	S	.282	.252[a]	.252[a]	.250	.237[a]	.237
	M	.574	.514[a]	.514[a]	.552	.508[a]	.508[a]
	L	.915	.902	.902	.919	.899[a]	.899
(3,5)	S	.261	.253[a]	.253[a]	.262	.236[a]	.236[a]
	M	.561	.506[a]	.506[a]	.557	.497[a]	.497[a]
	L	.918	.877[a]	.877[a]	.923	.890[a]	.890[a]
(4,5)	S	.272	.259	.253	.286	.262	.263[a]
	M	.567	.535[a]	.540	.563	.523[a]	.528[a]
	L	.922	.902[a]	.909	.920	.886[a]	.905
(5,5)	S	.292	.306	.297	.246	.233[a]	.243
	M	.610	.589	.604	.553	.512[a]	.541
	L	.919	.883[a]	.913	.915	.874[a]	.904

[a]The observed proportion is more than 2 σ_p from the observed proportion for t, where, for S, M, and L, σ_p = .0145, .0155, and .0095, respectively.

TABLE 2 (Continued)

Skewed Population

Sample Size	θ	Small Sample Addition Procedure			Large Sample Addition Procedure		
		t	U	Permutation t	t	U	Permutation t
(2,3)	S	.324	.313	.313[a]	.309	.354[a]	.354[a]
	M	.676	.644[a]	.644[a]	.686	.690	.690
	L	.885	.883	.883	.917	.917	.917
(3,3)	S	.399	.405	.405	.368	.370	.370
	M	.680	.683	.683	.659	.657	.657
	L	.893	.897	.897	.895	.879	.879
(2,4)	S	.378	.327[a]	.327[a]	.368	.442[a]	.442[a]
	M	.658	.622[a]	.622[a]	.697	.738[a]	.738[a]
	L	.894	.860[a]	.860[a]	.913	.914	.914
(3,4)	S	.377	.368	.368	.394	.432	.432
	M	.664	.645	.645	.693	.698	.698
	L	.898	.866[a]	.866[a]	.890	.885	.885
(4,4)	S	.389	.406	.406	.396	.384	.384
	M	.666	.632[a]	.632[a]	.673	.661	.661
	L	.899	.848[a]	.848[a]	.899	.866[a]	.866[a]
(2,5)	S	.358	.280[a]	.280[a]	.353	.460[a]	.460[a]
	M	.650	.563[a]	.563[a]	.703	.739[a]	.739[a]
	L	.881	.844[a]	.844[a]	.906	.912	.912
(3,5)	S	.379	.346[a]	.346[a]	.391	.444[a]	.444[a]
	M	.672	.622[a]	.622[a]	.686	.668	.668
	L	.890	.851[a]	.851[a]	.902	.877[a]	.877[a]
(4,5)	S	.382	.373	.380	.357	.365	.365
	M	.672	.634[a]	.647	.628	.598	.621
	L	.885	.844[a]	.881	.868	.812	.856
(5,5)	S	.331	.380[a]	.348	.356	.388[a]	.367
	M	.638	.646	.645	.661	.667	.672
	L	.889	.862[a]	.884	.892	.860[a]	.886

[a]The observed proportion is more than $2\sigma_p$ from the observed proportion for t, where, for S, M, and L, σ_p = .0145, .0155, and .0095, respectively.

than expected by chance with large differences occurring for the uniform distribution. Of the six cases where the power of the permutation t-test was larger than that of Student's t-test, two occurred for sample size (5,5). Also, for sample sizes (4,5) and (5,5) the power values of the permutation t-test were generally closer to those of Student's t-test than was true for the smaller sample sizes. For sample sizes (4,5) and (5,5) the values of empirical power of the permutation t-test were usually larger than those of the Mann-Whitney U-test.

For the "large sample addition procedure" when sampling from the skewed population, the empirical power values for the permutation t-test were greater than those of Student's t-test for fifteen of the twenty-seven cases available (three levels of θ, nine sample sizes). The seven differences which were larger than expected from sampling variation were for unequal sample sizes with either the small or medium levels of power. For example, for sample size (2,3), with small θ, the values .354 for the permutation t-test and .309 for Student's t-test are more than $2\sigma_p$ apart and thus are most likely due to something other than sampling variation. Other large differences occurred for sample size (3,5) with small and medium θ. For samples of equal size < (3,3), (4,4), (5,5) < or near equal size < (3,4), (4,5) < the differences between the power values for Student's t-test and the permutation t-test were small.

For the "small sample addition procedure" when sampling from the skewed population, or when the smaller sample came from the skewed population with the larger mean ($\mu + \theta$), the empirical power values for the

permutation t-test were greater than or equal to those for Student's t-test for only six of the twenty-seven comparisons available. In fact thirteen of the twenty-seven comparisons showed a larger-than-sampling-variation difference with Student's t-test having the larger power value. The differences in favor of Student's t-test were the largest for unequal sample sizes, and only for equal sample sizes (3,3) and (5,5) did the power values of the permutation t-test approach or exceed those of Student's t-test.

IV

SUMMARY AND CONCLUSIONS

The results presented above for the permutation t-test show that the empirical probability of a Type I error for repeated sampling from a normal or uniform population was generally close to the theoretical α. The empirical probability of a Type I error for repeated sampling from the specified skewed population was generally close to the theoretical α but showed one sample size which had inexplicably divergent results for the permutation t-test and the Mann-Whitney U-test. This discrepancy was part of a trend of other discrepancies which were within the bounds of sampling variation. The empirical results for the power showed that the permutation t-test generally had smaller power than Student's t-test for the uniform and normal populations. For the skewed population the permutation t-test generally had higher power values than Student's t-test if the larger sample were drawn from the population with the larger mean $(\mu + \theta)$. If the samples were of equal size, the permutation t-test generally had power values which were close to those of Student's t-test but did not exceed them. However, if the smaller sample were drawn from the skewed population with the larger mean $(\mu + \theta)$, then the power values of Student's t-test were larger than those of the permutation t-test. For all three populations, the power of the permutation t-test approached that of Student's t-test as sample size increased, even for samples as small as (4,5) and (5,5). The increase in power was more rapid for the permutation t-test than for the Mann-Whitney U-test, and the power of the permutation t-test was always greater than or equal to that of the Mann-Whitney U-test.

Using Student's t as the test statistic for the permutation test for the two-sample problems gives a statistical procedure which not only has ARE of one for the normal population but has very close agreement with Student's t-test for small samples. The agreement is indicated by the closeness of values of empirical power and probability of a Type I error for the permutation t-test when compared to those of Student's t-test for the normal population. Although similar results show that the permutation t-test is in close agreement with Student's t-test for the uniform population, the empirical power of the permutation t-test for the skewed population showed that the permutation t-test could have higher power than Student's t-test if the sample sizes were proportional to the population means when the parent population has the specific skewed distribution with $\gamma_1 = 1.633$ and $\gamma_2 = 4$. The present study also gives further support to the knowledge that Student's t-test is generally robust to the violation of the normality assumption, even for very small samples.

The present research indicates that the permutation t-test is an acceptable statistical procedure for the two-sample problem for the normal and uniform populations and suggests that it might be more desirable than the traditional Student's t-test when sample sizes are proportional to the means and the parent population is nonnormal and asymmetric. Further research is needed before a more definite statement can be made about the permutation t-test when sampling from nonnormal populations.

Appendix A

An Example on Permutations and Combinations

For example, consider m=n=2 and ΣX as the statistic:

Permutations

X	Y	ΣX
12	34	3
12	43	3
21	34	3
21	43	3
13	24	4
13	42	4
31	24	4
31	42	4
14	23	5
14	32	5
41	23	5
41	32	5
23	14	5
23	41	5
32	14	5
32	41	5
24	13	6
24	31	6
42	13	6
42	31	6
34	12	7
34	21	7
43	12	7
43	21	7

Combinations

X	Y	ΣX
12	34	3
13	24	4
14	23	5
23	14	5
24	13	6
34	12	7

For Both Permutations and Combinations

ΣX	$p(\Sigma X)$
3	1/6
4	1/6
5	2/6
6	1/6
7	1/6

Appendix B

Values of θ for Various Sample Sizes

Sample Sizes	Small θ	Medium θ	Large θ
(2,3)	.7254	1.6327	2.7726
(3,3)	1.0888	1.8783	2.9390
(2,4)	.9631	1.7790	2.8554
(3,4)	1.2758	2.0195	3.0243
(4,4)	1.1418	1.8011	2.6854
(2,5)	1.1026	1.8765	2.9084
(3,5)	1.0745	1.7427	2.6354
(4,5)	1.0145	1.6175	2.4214
(5,5)	.7871	1.3336	2.0546

REFERENCES

Baker, F. B., & Collier, R. O. (1966a). An empirical study into factors affecting the F-test under permutation for the randomized block design. J. Amer. Statist. Ass.

Baker, F. B., & Collier, R. O. (1966b). Monte Carlo F-II: A computer program for analysis of variance F-tests by means of permutation. Educational and Psychological Measurement, 26, No. 1.

Box, G. E. P. (1954a). Some theorems on quadratic forms applied in the study of analysis of variance problems: I. Effect of inequality of variance in a one-way classification. Ann. Math. Statist., 25, 290-302.

Box, G. E. P. (1954b). Some theorems on quadratic forms applied in the study of analysis of variance problems: II. Effects of inequality of variance and covariance between errors in the two-way classification. Ann. Math. Statist., 25, 484-498.

Box, G. E. P., & Andersen, S. L. (1955). Permutation theory in the derivation of robust criteria and the study of departures from assumptions. J. Roy. Statist. Soc. (B), 17, 1-34.

Collier, R. O., & Baker, F. B. (1963). The randomization distribution of F-ratios for the split-plot design--an empirical investigation. Biometrika, 50, 431-438.

Collier, R. O., & Baker, F. B. (1966). Some Monte Carlo results on the power of the F-test under permutation in the simple randomized block design. Biometrika, 53, 199-203.

Control Data Corporation (1966). 3600-3800 Computer Systems. Library Functions, p. 89.

Dixon, W. J. (1954). Power under normality of several non-parametric tests. Ann. Math. Statist., 25, 610-614.

Festinger, L. (1946). The significance of differences between means without reference to the frequency of distribution function. Psychometrika, 11, 97-105.

Fisher, R. A. (1925). Statistical Methods for Research Workers. New York: Hafner Publishing Company, Inc.

Fisher, R. A. (1935). The Design of Experiments. Edinburgh: Oliver and Boyd.

Gayen, A. K. (1949). The distribution of "Student's" t in random samples of any size drawn from non-normal universes. Biometrika, 36, 353.

Gayen, A. K. (1950). The distribution of the variance ratio in random samples of any size drawn from non-normal universes. Significance of difference between the means of two non-normal samples. Biometrika, 37, 236-399.

Geisser, S., & Greenhouse, S. W. (1958). An extension of Box's results on the use of the F-distribution in multivariate analysis. Ann. Math. Statist., 29, 885-891.

Gibbons, J. D. (1964). On the power of two-sample rank tests on the equality of two distribution functions. J. Roy. Statist. Soc. (B), 26, 293-304.

Glazer, H. (1964). Comparison of the Sutdent two-sample t-test and the Wilcoxon-Mann-Whitney test for normal distributions with unequal variances. Unpublished doctoral dissertation, Boston University.

Haynam, G. E., & Govindarajulu, Z. (1966). The exact power of the Mann-Whitney test for exponential and rectangular alternatives. Ann. Math. Statist., 36, 945-953.

Hays, W. L. (1965). Statistics for Psychologists. New York: Holt, Rinehart and Winston.

Hodges, J. L., Jr., & Lehmann, E. L. (1956). The efficiency of some non-parametric competitors of the t-test. Ann. Math. Statist., 27, 324-335.

Hoeffding, W. (1952). The large sample power of tests based on permutations of observations. Ann. Math. Statist., 23, 169-192.

Hotelling, H., & Pabst, M. R. (1936). Rank correlation and tests of significance involving no assumptions of normality. Ann. Math. Statist., 7, 29-43.

Hull, T. E., & Dobell, A. R. (1962). Random number generators. SIAM Review, 4, No. 3, 230-254.

International Business Machines Corporation (1959). Random number generation and testing. Reference manual c20-8011.

Kempthorne, O. (1952). The Design and Analysis of Experiments. New York: John Wiley and Sons, Inc.

Kempthorne, O., Zyskind, G., Addleman, S., Throckmorton, T., & White, R. (1961). Analysis of variance procedures. ARL Report 144, U.S.A.F.

Kendall, M. G., & Stuart, A. (1967). The Advanced Theory of Statistics, Vol. 2. London: Charles Griffin and Co., Ltd.

Kolmogorov, A. N. (1941). Confidence limits for an unknown distribution function. Ann. Math. Statist., 12, 461-463.

Kruskal, W. H., & Wallis, W. A. (1952). Use of ranks on one-criterion variance analysis. J. Amer. Statist. Ass., 47, 583-621.

Lehmann, E. L. (1953). The power of rank tests. Ann. Math. Statist., 24, 23-43.

Mann, H. B., & Whitney, D. R. (1947). On a test of whether one of two random variables is stochastically larger than the other. Ann. Math. Statist., 18, 50-60.

Milton, R. C. (1964). An extended table of critical values for the Mann-Whitney (Wilcoxon) two-sample statistic. J. Amer. Statist. Ass., 59, 925-934.

Milton, R. C. (1966). Rank order probabilities: Two-sample normal shift alternatives. Technical Report No. 53a, University of Minnesota Department of Statistics.

Mood, A. M. (1950). Introduction to the Theory of Statistics. New York: McGraw-Hill Book Company.

Mood, A. M., & Graybill, F. A. (1963). Introduction to the Theory of Statistics. New York: McGraw-Hill Book Company.

Pearson, K. (1911). On the probability that two independent distributions of frequency are really samples from the same populations. Biometrika, 8, 250-254.

Pitman, E. J. G. (1937a). Significance tests which may be applied to samples from any populations. J. Roy. Statist. Soc. (B), 4, 119-130.

Pitman, E. J. G. (1937b). Significance tests which may be applied to samples from any population: III. The analysis of variance test. Biometrika, 29, 322-335.

Pratt, J. W. (1964). Robustness of some procedures for the two-sample location problem. J. Amer. Statist. Ass., 59, 665.

Savage, I. R. (1962). Bibliography of Nonparametric Statistics. Cambridge: Harvard University Press.

Scheffé, H. (1943). Statistical inference in the nonparametric case. Ann. Math. Statist., 14, 305-332.

Scheffé, H. (1959). The Analysis of Variance. New York: John Wiley and Sons, Inc.

Siegel, S. (1956). Nonparametric Statistics for the Behavioral Sciences. New York: McGraw-Hill Book Co.

Smirnov, N. V. (1948). Tables for estimating the goodness of fit of empirical distributions. Ann. Math. Statist., 19, 279-281.

Srivastava, A. B. L. (1958). Effect of non-normality on the power function of the t-test. Biometrika, 45, 421.

Student (1908). The probable error of a mean. Biometrika, 6, 1.

Toothaker, L. E. (1967). An empirical investigation of the combined effect of heterogeneity of variance and the existence of block-treatment interaction on the F-test under permutation for the randomized block design. Unpublished Master of Science thesis, University of Wisconsin.

University of Wisconsin Computing Center (UWCC). User's Manual, Vol. IV, Sec. 3.13, RANSS.

van der Vaart, H. R. (1950). Some remarks on the power function of Wilcoxon's test for the problem of two samples I, II. Proc. Koninkl. Ned. Akad. v. Wet., A, 53, 494-520.

van der Vaart, H. R. (1961). On the robustness of Wilcoxon's two-sample test. Quantitative Methods in Pharmacology, H. de Jonge, Ed. New York: Interscience, pp. 140-158.

Wald, A., & Wolfowitz, J. (1940). On a test whether two samples are from the same population. Ann. Math. Statist., 11, 147-162.

Wald, A., & Wolfowitz, J. (1943). An exact test for randomness in the non-parametric case based on serial correlation. Ann. Math. Statist., 14, 378-388.

Wald, A., & Wolfowitz, J. (1944). Statistical tests based on the permutations of the observations. Ann. Math. Statist., 15, 358-372.

Welch, B. L. (1937). On the z-test in randomized blocks and Latin squares. Biometrika, 29, 21-52.

Wetherill, G. B. (1960). The Wilcoxon test and non-null hypothesis. J. Roy. Statist. Soc. (B), 22, 402-418.

Wilcoxon, F. (1945). Individual comparison by ranking methods. Biometrics Bull., 1, 80-83.

Wilks, S. S. (1962). Mathematical Statistics. New York: John Wiley and Sons, Inc.

GPO 826–076